Model-Based Clustering for Image Segmentation and Large Datasets Via Sampling[1]

Ron Wehrens and Lutgarde M.C. Buydens
University of Nijmegen

Chris Fraley and Adrian E. Raftery
University of Washington

Technical Report no. 424
Department of Statistics
University of Washington.

February 13, 2003

[1]Ron Wehrens is associate professor and Lutgarde M.C. Buydens is professor, both at the Department of Analytical Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, Netherlands. Email: rwehrens|lbuydens@sci.kun.nl; web: www.sci.kun.nl/cac. Chris Fraley is a research staff member and Adrian E. Raftery is Professor of Statistics and Sociology, both at the Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322; Email: fraley|raftery@stat.washington.edu; Web: www.stat.washington.edu/fraley|raftery. The research of Fraley and Raftery was supported by NIH grant 1R01CA094212-01, and by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
13 FEB 2003		**00-02-2003 to 00-02-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Model-Based Clustering for Image Segmentation and Large Datasets Via Sampling	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**26**	

Abstract

The rapid increase in the size of data sets makes clustering all the more important to capture and summarize the information, at the same time making clustering more difficult to accomplish. If model-based clustering is applied directly to a large data set, it can be too slow for practical application. A simple and common approach is to first cluster a random sample of moderate size, and then use the clustering model found in this way to classify the remainder of the objects. We show that, in its simplest form, this method may lead to unstable results. Our experiments suggest that a stable method with better performance can be obtained with two straightforward modifications to the simple sampling method: several tentative models are identified from the sample instead of just one, and several EM steps are used rather than just one E step to classify the full data set. We find that there are significant gains from increasing the size of the sample up to about 2,000, but not from further increases. These conclusions are based on the application of several alternative strategies to the segmentation of three different multispectral images, and to several simulated data sets.

Keywords: EM algorithm; MRI image; Remote sensing; Sampling

Contents

List of Tables

List of Figures

1 Introduction

Today, data are generated at unprecedented speed. The growing size of data sets and data bases has increased the need for good clustering methods to capture and summarize the information. An example is the segmentation of multispectral images, where the objective is to group similar pixels, and to assess how many different groups there are. Typically, three to ten congruent images or bands containing complementary information are recorded, often containing tens of thousands of pixels per image. This places constraints on clustering techniques with respect to memory usage and computing time.

Many different clustering methods have been described (Jain and Dubes 1988; Kaufman and Rousseeuw 1989). Model-based clustering (McLachlan and Basford 1988; Banfield and Raftery 1993; Fraley and Raftery 2002b; McLachlan and Peel 2000) is one of the more recent developments, and has shown very good performance in a number of fields (Mukherjee, Feigelson, Babu, Murtagh, Fraley, and Raftery 1998; Dasgupta and Raftery 1998; Yeung, Fraley, Murua, Raftery, and Ruzzo 2001; Wang and Raftery 2002), including image analysis (Campbell, Fraley, Murtagh, and Raftery 1997; Campbell, Fraley, Stanford, Murtagh, and Raftery 1999; Stanford and Raftery 2002; Wehrens, Simonetti, and Buydens 2002). As implemented in these applications, and in available software (**?**; McLachlan, Peel, Basford, and Adams 1999; Fraley and Raftery 2002a), model-based clustering consists of fitting a mixture of multivariate normal distributions to a data set by maximum likelihood using the EM algorithm, possibly with geometric constraints on the covariances matrices, and an additional component to allow for outliers or noise. Since the likelihood surface typically has many local maximna, initialization of the EM algorithm is a very important issue. Model-based hierarchical clustering (Banfield and Raftery 1993) has been found to provide good initializations.

Model-based hierarchical clustering generally requires storage and computing time at least proportional to the square of the dimension of the data, so that both space and time are limiting factors in its application to large data sets. Another problem is that when the size of the data set reaches a certain threshold, it is not possible to keep all of the required quantities in memory at the same time, forcing a dramatic and abrupt increase in necessary computational resources. This threshold varies with computer hardware and software, and data dimension, but at the current time it is typically on the order of several thousand objects.

Various approaches to the problem of clustering large data sets have been proposed, including initialization by clustering a sample of the data (Banfield and Raftery 1993; Fayyad and Smyth 1996; Maitra 2001), and using an initial crude partitioning of the entire data set (Posse 2001; Tantrum, Murua, and Stuetzle 2002). The simplest and perhaps most widely applied approach is to apply the clustering method first to a small simple random sample from the data, and then apply the resulting estimated model to the full data set using discriminant analysis (Banfield and Raftery 1993). The discriminant analysis can be carried out very easily in the model-based clustering framework by using a single E step (Fraley and Raftery 2002b).

Unfortunately, this easily implemented strategy may lead to unstable segmentations

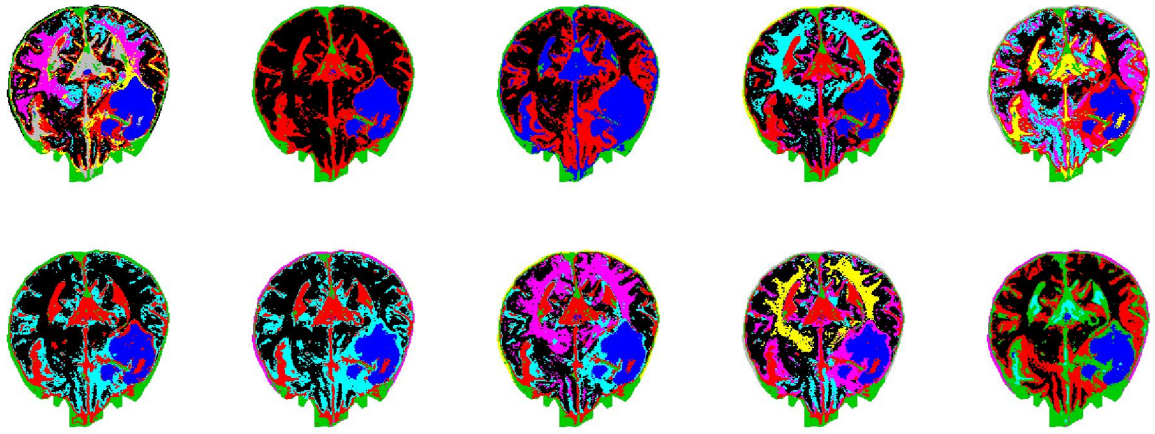

Figure 1: Segmentations based on clustering of samples of 500 pixels of a set of four congruent MRI images of a patient with a brain tumour. Colors are permuted to give maximal visual similarity.

when used in its simplest form, as illustrated in Figure 1. In this figure, ten random samples of 500 pixels are used to cluster an MRI data set of a patient with a brain tumour. The number of clusters, selected by the method (see below), varies between 4 and 9. Although several features are preserved, such as the tumour region (in dark blue, behind the eye) and the cerebrospinal fluid (green), the variation is quite large.

In this paper we show that the appealingly simple approach based on clustering a sample of the data can be modified to give good and stable results, with two straightforward changes. For the image data sets we consider, we obtain good results by tentatively selecting several models based on the sample rather than just one, and by running several EM steps on the full data set rather than just one E step. We find that performance improves when the size of the sample is increased up to about 2,000, but that beyond that there is little gain. To reach this conclusion, we considered a range of sample sizes and several strategies of varying computational cost. Comparisons were based on three typical real-world data sets, and several realistic simulations.

In Section 2, we give a brief overview of model-based clustering, and in Section 3 we propose several strategies for model-based clustering in large data sets such as those that arise in image segmentation. The image data we use and the design of our simulations is described in Section 4. Segmentations using different sample sizes and strategies are compared in Section 5 based on the likelihoods of the segmented images, the stability of the clusters, and the accuracy of the results in the simulated cases. Finally, recommendations are made.

2 Model-Based Clustering

In model-based clustering, individual clusters are described by multivariate normal distributions, where the class labels, parameters and proportions are unknown. Maximum likelihood estimates for the resulting model can be obtained using the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977; McLachlan and Krishnan 1997). Given an initial guess for the cluster means μ_k, covariances $\mathbf{\Sigma}_k$, and proportions τ_k for all clusters, one can calculate the conditional probability that object i belongs to cluster k:

$$z_{ik} = \tau_k \phi_k(\mathbf{x}_i|\mu_k, \mathbf{\Sigma}_k) / \sum_{j=1}^{K} \tau_j \phi_j(\mathbf{x}_i|\mu_j, \mathbf{\Sigma}_j, \tau_j) \ ,$$

where ϕ is the multivariate normal (Gaussian) density. This is the expectation step, or E-step of the EM algorithm. The maximization step (M-step) consists of estimating the parameters $\mu, \mathbf{\Sigma}$, and τ, from the data and the conditional probabilities z_{ik}. The E- and M-steps iterate until convergence. Finally, each object is classified in the class in which it has the highest conditional probability.

Good initialization of the EM algorithm is very important, since the method may converge to different values depending on where it is started because the likelihood surface usually has multiple local maxima. For the initialization, we apply fast hierarchical model-based clustering (Fraley 1998), the default in the MCLUST software (**?**; Fraley and Raftery 2002a).

If there are no cross-cluster constraints on the cluster shapes and sizes, each one is described by $1 + p + p(p+1)/2$ parameters (the proportion, mean and covariance matrix, respectively). The covariance matrix for the kth cluster can be expressed in the form

$$\mathbf{\Sigma}_k = \lambda_k D_k A_k D_k^T \tag{1}$$

where λ_k describes the volume of the cluster, D_k is the matrix of eigenvectors, governing the orientation of the cluster, and A_k is a diagonal matrix, proportional to the eigenvalues, which determines the shape of the cluster. Banfield and Raftery (1993) proposed cross-cluster equality constraints on any or all of the cluster volumes, orientations or shapes based on this decomposition as a way of limiting the number of parameters in the model in a geometrically intuitive way. One such model constrains all clusters to have the same shape, but allows cluster volumes and orientation to vary. This is called the VEV model (Fraley and Raftery 1999) (Variable volume, Equal shape, Variable orientation). A completely unconstrained model is denoted by VVV (Fraley and Raftery 1999). For a discussion of all possible combinations of constraints based on the decomposition (1), see Celeux and Govaert (1995). Each set of constraints corresponds to a different clustering criterion: for example, if the clusters are restricted to be spherical and identical in volume, the criterion is the same as that used in Ward's clustering and standard k-means clustering (Celeux and Govaert 1995; Fraley and Raftery 1998).

The model-based clustering framework can be extended in a natural way to model noise and outliers (Banfield and Raftery 1993; Fraley and Raftery 2002b): an extra "noise" class

is described by a constant component density over the whole data region. If the densities of all other components are lower than the density of the noise class, an object will be classified as noise. This corresponds to modelling the noise with a homogeneous Poisson process. An initial estimate of the noise is needed.

To select the optimal clustering model (defined by both the cross-cluster constraints and the number of clusters), several measures have been proposed (for an overview, see e.g. McLachlan and Peel 2000). In several applications, the BIC approximation to the Bayes factor (Schwarz 1978; Kass and Raftery 1995) has performed quite well (Fraley and Raftery 1998; Dasgupta and Raftery 1998; Stanford and Raftery 2000). The strategy employed here thus consists of several steps (Fraley and Raftery 1998): first perform model-based hierarchical clustering for initialization; then perform EM for several values of the number of clusters and with several sets of constraints on the covariance matrices of the clusters; finally, select the combination of model and number of groups that leads to the highest BIC value.

The model-based clustering framework also provides a measure of the certainty of a particular classification. Basically, a classification has low uncertainty if one of the k conditional cluster membership probabilities for each data point is close to 1, and the other $(k-1)$ conditional probabilities ar close to 0. Bensmail et al. (1997) quantified this notion by defining the uncertainty of the classification of object i to be

$$u_i = 1 - \max_k z_{ik}.$$

The uncertainty of the complete clustering may then be estimated by averaging over the uncertainty of all objects.

For large data sets, the usual strategy is to apply model-based clustering to a random sample from the data set of a size that can be clustered comfortably (Banfield and Raftery 1993; Fraley and Raftery 2002b). The parameters of the clusters, found in the sample, can be used to classify the remainder of the objects by the application of a single E-step for the entire data set, which is very quick.

3 Sampling Methods

The segmentations in Figure 1 were obtained by clustering ten random samples of 500 pixels each to obtain ten sets of cluster parameters, and for each random sample performing one E-step to classify all other pixels into one of the clusters. The variability that is so apparent can have several causes: first, the sample size may be too small, so that clusters are not well described. In particular, there is a danger that one misses small clusters with sample sizes that are too small. To investigate the effect of sample size, five different sample sizes are compared: 500, 1000, 1500, 2000 and 2500 pixels, respectively. With current software and hardware, 2500 pixels can be clustered within a reasonable time.

Second, there may be a problem in going from the clustered sample to a complete segmented image. We will compare four strategies. The strategy described earlier, applying one E-step using the cluster parameters from the clustered sample, will be called strategy I.

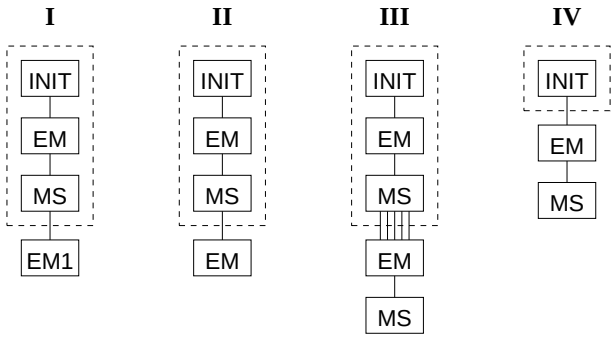

Figure 2: Strategies to apply model-based clustering for large data sets. The dashed box contains operations on the sample only. INIT: initialization by model-based hierarchical clustering; EM: application of the EM algorithm to find cluster parameters and classifications (max. 100 steps); MS: model selection; EM1: one iteration of the EM algorithm to classify pixels not in the sample.

It may be beneficial to do several EM steps on the complete image; this is feasible in terms of computational cost. The second strategy (II) extends strategy I by doing an additional EM optimization for at most 100 steps for the selected model, taking into account all pixels. The sample, however, may be too small to pick the correct model. The third strategy (III) therefore does at most 100 EM steps for the five best models, selected on the basis of the training set, and from these five eventually selects the best using the whole data set. Finally, the fourth strategy (IV) uses the sample only for the initialization, and does 100 EM steps for all models considered. Only then is the best model selected. Strategy IV can be viewed as a gold standard, but it is much more computationally expensive than the other ones. The four strategies are summarized in Figure 2.

Each experiment is performed ten times with different random samples; strategies I-IV were implemented in exactly the same way for each experiment.

4 Data and Simulations

4.1 Assessment of Results

Several criteria are used to assess the effects of the sample size and the strategy. One aspect is the stability of the clustering: ideally, the results should be independent of the initial sample. This means that the models selected in the ten repeated experiments should be similar, with, ideally, the same model being selected in most or all of the experiments. It also means that, even if there are differences between the selected models, the final classifications should be similar. The latter is assessed by calculating the adjusted Rand index (Rand 1971; Hubert 1985). One can also consider the likelihoods of the final segmentations. Some samples may lead to local maxima which may be easily recognized. This gives an indication of what fraction should be used in the training phase and what strategy is

Figure 3: The three data sets, plotted in order of increasing size. The T1-weighted MRI image of a patient with a brain cancer behind the left eye (left), an image of a St. Paulia (middle), and a false-color image of the remote sensing data of the Duursche Waarden (right).

best.

A second aspect is the accuracy of the segmentation: how similar are the estimated clusters to the "true" clusters? We use simulated data to assess this accuracy, which depends on sample size and clustering strategy. Since we know the "correct" model, we can count the number of times the correct model is picked under the four strategies and with the varying sample sizes. Also, we can simply count the errors in the classifications, since the "true" classification is known. Again, the adjusted Rand index is used to quantify the agreement between cluster labels and "true" labels.

4.2 Data

Three real data sets will be used. The first is a set of four congruent MRI images (T1-weighted, T2-weighted, proton-density and gadolineum-enhanced) of a patient with a brain tumour. In this data set, uninteresting regions (eyes, the skull and the region outside the head) have been removed, leaving 23712 pixels. The T1-weighted image is depicted in the left plot of Figure 3. The second data set is an RGB image of a St. Paulia flower with 268 columns and 304 rows; again, pixels from the background have been removed so that the data set has 45656 pixels. The RGB image is plotted in Figure 3 (middle plot). The final data set (RS) is a 256 by 256 remote sensing image of an area in The Netherlands, the Duursche Waarden. It is recorded by an airborne CASI scanner, and consists of 9 spectral bands. A false-color image is shown in the right plot of Figure 3. (Bands 6, 3 and 1 are used for red, green and blue, respectively.) The vertical discontinuity, slightly left of the center, is due to the fusion of two flight lines.

4.3 Simulation Design

Four data sets were simulated by randomly drawing from a series of multivariate normal distributions obtained from clustering the MRI images. The first pair of two simulated data

sets (Simul12 and Simul12N) was based on the model that led to the highest full-image segmentation loglikelihood in strategy I (N=2500): (VEV,12). This is the VEV model with 12 clusters. From the model parameters, 30.000 points were generated randomly (in four dimensions, like the MRI data set). The Simul12N data set was derived from the Simul12 data set by replacing five percent of the pixels by uniformly distributed noise.

The second pair of simulated data sets (Simul6 and Simul6N) was based on a model with fewer clusters, in this case the (VEV,6) model (MRI, N=500, sample 7). Again, in the noise data set (Simul6N) five percent of the original pixels were replaced by uniformly distributed noise. In modelling the noise case, our initial estimate of the noise is based on the "true" noise, so that this is the best possible initialization. In practice, an imperfect initial estimate of the noise is likely to lead to a decrease in accuracy.

4.4 Software and Hardware

All experiments were performed in R (Ihaka and Gentleman 1996) version 1.6.0, using the 2002 version of MCLUST by Fraley and Raftery (Fraley and Raftery 1999; Fraley and Raftery 2002a). MCLUST considers ten parametrizations of the cluster covariance matrices (two spherical models, four diagonal models and four ellipsoidal models) (Fraley and Raftery 2002a). Adjusted Rand indices and associated plots are programmed in R. Scripts for the application of strategies I–IV are available as supplementary material (`www.sci.kun.nl/cac/people/rwehrens/software`).

We used an i686 (Pentium III, 1.0 GHz) computer, running RedHat Linux (kernel version 2.4.18-4smp). The stability of the EM calculations was checked by performing EM runs (maximally 100 steps) on a data matrix with permuted rows (RS data, so 65,536 rows), starting from models initialised on ten random samples of 500 pixels each. In all cases, differences between loglikelihoods of the unpermuted and permuted data were on the order of 10^{-8}, so there seem to be no stability problems in the EM steps. For strategies II and III, less than 100 EM steps were needed to reach convergence in all cases; for strategy IV, this was the case for all relevant models. Occasionally, inappropriate models did not converge within 100 steps.

5 Results

For all data sets, we considered clusterings with 1 to 20 clusters, using strategies I–III. Because of time constraints, fewer possibilities were considered for strategy IV: in the MRI case 4–18 clusters, for the RS data set 4–11 clusters, and for the St. Paulia image 3–15 clusters. For the smaller data sets, all ten model parametrizations available in MCLUST were considered in strategies I–III; for the RS data set, only the four most elaborate models were considered (EEE, EEV, VEV and VVV). For strategy IV only the EEV, VEV and VVV models were considered.

Table 1: Most frequently selected models in strategy I. The columns indicate different sample sizes. Numbers in brackets indicate how often a particular model was selected. The true model for Simul12 and Simul12N is VEV12; the true model for Simul6 and Simul6N is VEV6. Simul12N and Simul6N contain 1500 noise points (5 percent).

Sample size	500	1000	1500	2000	2500
MRI	VEV6,7,9 (2)	VEV7 (5)	VEV10 (3)	VEV7 (3)	VEV8 (3)
Paulia	VEV4 (3)	VEV7 (3)	VEV7,9 (3)	VEV8 (4)	VEV9 (4)
RS	VVV4 (8)	VVV6 (5)	VVV6 (9)	VVV6,7 (5)	VVV9 (4)
Simul12	VEV4 (4)	VEV8,9 (4)	VEV7,8 (3)	VEV10–13 (2)	VEV12 (5)
Simul12N	VEV4 (3)	VVV5, VEV7,8(2)	VEV9 (5)	VEV9 (4)	VVV8, VEV9 (3)
Simul6	VEV6 (5)	VEV6 (5)	VEV6 (7)	VEV6 (6)	VEV6 (8)
Simul6N	VEV5 (4)	VEV6 (4)	VEV6,7 (4)	VEV6 (6)	VEV7 (6)

5.1 Stability of Model Selection

The images in Figure 1, obtained by applying strategy I, with a sample size of 500, correspond to seven different clustering models, among which the (VEV,6), (VEV,7) and (VEV,9) models occur twice. Results like this are summarized in Table 1. Sample sizes range from 500 to 2500.

For all images, the complexity of the selected model (i.e. the number of clusters) is observed to increase with the sample size. Apparently, more clusters are needed to describe the sample. This suggests either that some smaller clusters are missed with the smallest samples, or that the data are not exactly normally distributed and that including more Gaussian components in the mixture leads to a better fit. The effect is clearest for the RS data but is also found in the MRI and St. Paulia images, although it is not so clear from Table 1: the variability in the selected models is much larger than with the RS image. In the case of a sample of 500 pixels from the St. Paulia image, models with 3–6 clusters were selected; in the case of samples of 2500 pixels, the models selected had 7–15 clusters. The true model for the simulated data sets Simul12 and Simul12N is known to be (VEV,12), and models close to this are selected only for samples of at least 2000 pixels. On the other hand, all strategies and all sample sizes are able to pick the correct model (or a close one) for the VEV,6 models of the Simul6 and Simul6N data sets. Model selection for strategy II is the same as for strategy I, so we do not consider strategy II further in this section.

Strategy III consists of doing EM on the whole image for the five models with the highest BIC values for the sample. This invariably leads to more complex models than strategy I. If a VVV model is selected in strategy I, strategy III will typically select a VVV model with one or two extra clusters or, less often, a VEV model with 5-6 extra clusters; if a VEV model is chosen by strategy I, strategy III will pick a VVV model with the same number of clusters, or with one or two extra clusters, or a VEV model with more clusters (see Table 2).

The general trends, however, do not change: a larger sample leads to a more complex cluster model, and the variability in the models is much smaller for the RS image than for the other two images. If anything, the variability in the selected models is larger in

Table 2: The most frequently selected models in strategy III. See caption of Table 1.

Sample size	500	1000	1500	2000	2500
MRI	VEV9 (4)	VEV10 (5)	VEV11 (3)	VEV13, VVV8 (3)	VEV10,12,13, VVV10 (2)
St. Paulia	VEV6 (4)	VEV10 (3)	VVV8 (3)	VVV9 (2)	VEV14 (4)
RS	VVV5 (7)	VVV8 (8)	VVV9 (8)	VVV10 (5)	VVV10,11 (5)
Simul12	VEV6 (4)	VEV10 (5)	VEV11,13 (3)	VEV12,13 (3)	VEV12,14 (3)
Simul12N	VEV6 (6)	VEV9 (5)	VEV10 (5)	VEV11(4)	VEV9, VEV11,12 (2)
Simul6	VVV6 (4)	VVV6 (5)	VVV6 (7)	VVV6 (5)	VVV6 (8)
Simul6N	VVV6 (3)	VVV6 (3)	VVV7,8 (3)	VVV6(4)	VVV7 (4)

Table 3: The most frequently selected models in strategy IV. The range of models considered is much smaller than with the other strategies (see text).

N	500	1000	1500	2000	2500
MRI	VVV18 (6)	VVV18 (8)	VVV18 (7)	VVV18 (8)	VVV18 (5)
St. Paulia	VVV15 (6)	–	–	–	–
RS	VEV15 (5)	–	–	–	–
Simul12	VEV13,14 (3)	VEV13 (4)	VEV15 (5)	VEV13 (5)	VEV13 (3)
Simul6	VEV6 (7)	VEV7 (5)	VEV6 (7)	VEV6 (7)	VEV6 (8)

strategy III than in strategy I. For the simulated data with 12 clusters, small samples lead to an underestimation of the complexity of the model. For samples of 1000 pixels and more, the models selected are approximately correct; this is an improvement compared to strategy I. The simulated data with six clusters seem to be overfit slightly: instead of a VEV model a VVV model is usually selected.

Strategy IV is the most computationally expensive one; the sample is used only to initialize the clustering, and EM is performed using the full data set for all models. The results, using a limited set of models and a restricted range of numbers of clusters, are summarized in Table 3. For the Simul6 data set the correct model is retrieved in most of the runs; the model selected for the Simul12 data set is usually slightly more complex than the model that generated the data. In the real data sets, in almost all cases the most complex cluster model possible is selected.

5.2 Stability of Segmentations

More important than the actual models selected is the question of how different the segmentation of the complete image is for the different samples. To assess this, for each data set the clusterings from the ten samples are compared with the adjusted Rand indices; the result is the mean value of all possible 45 comparisons. To present an intuitive calibration of the scale of the adjusted Rand index, a graphical impression is presented in Figure 4. The comparisons are between the first and second, third and second and seventh and sixth segmented images of Figure 1, respectively. The adjusted Rand index of 0.73 for the last comparison is the highest found in this set of segmented images. One can see an obvious

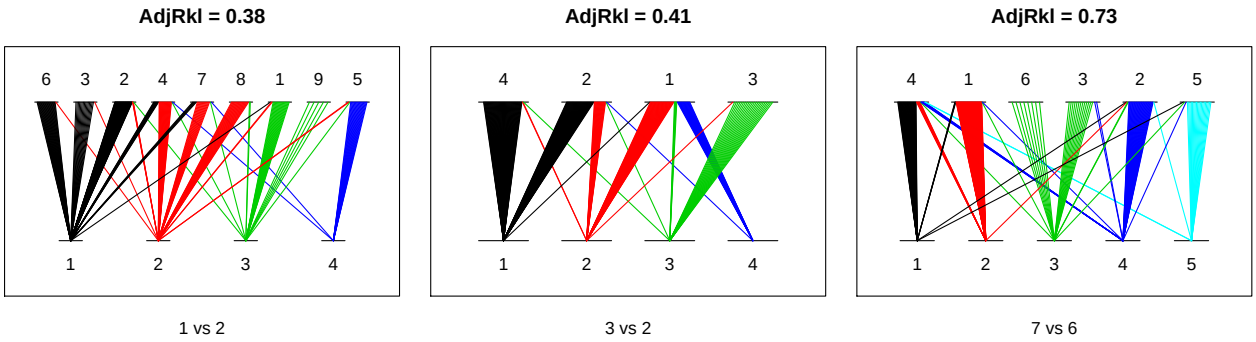

Figure 4: Adjusted Rand indices for the comparison of segmented images from Figure 1: 1 vs. 2; 3 vs. 2; and 7 vs. 6. Colors are based on the segmentation of images 2, 2 and 6, respectively. Each line corresponds to at most 100 pixels.

disadvantage of the adjusted Rand index: when the number of clusters differs a great deal, as in the first example, the index will often be quite low. However, there is a good correspondence between several clusters in the classification of sample 1 with individual clusters in the segmentation based on sample 2. In comparing segmentation 3 with 2, one clearly finds mixed clusters, which leads to an adjusted Rand index more or less equal to the first comparison.

Means and standard deviations of adjusted Rand indices for the different strategies are shown in Figure 5. For the real data sets, mean agreements based on raw classifications are better for strategy II. Strategy III shows levels of agreement similar to or less than those for Strategy I. This is caused by the more complex models that are selected and the larger variability in selected models. In the MRI data, strategy IV even leads to the lowest agreements of all the strategies.

For the simulated data, strategy I is again the worst, but now the differences between the other strategies are negligible. Again, sample size does not seem to make much difference. Even with perfectly Gaussian data, there will be variability in the eventual segmentation with all three strategies considered.

To investigate whether "uncertain" classifications are more likely to be variable, we compared the clusterings of the ten replicate segmentations in each experiment as a function of the pixel uncertainty. For a set of five thresholds (0.05, 0.1, 0.2, 0.35 and 0.5, respectively) we only included those pixels with uncertainties smaller than the threshold and calculated the agreement between the clusterings. The results are shown in Figures 6 and 7, as mean adjusted Rand indices. In general, agreements increase when taking only "certain" classifications into account. This means that the clusters are located at approximately the same positions, and that the largest differences between repeated clusterings are at the edges, as may be expected.

However, when eliminating uncertain pixels, a large part of the image remains unclassified; for the threshold of 0.05, the proportions of pixels eliminated in the MRI, St. Paulia and RS images are typically close to 70, 60 and 90 percent, respectively. For the next

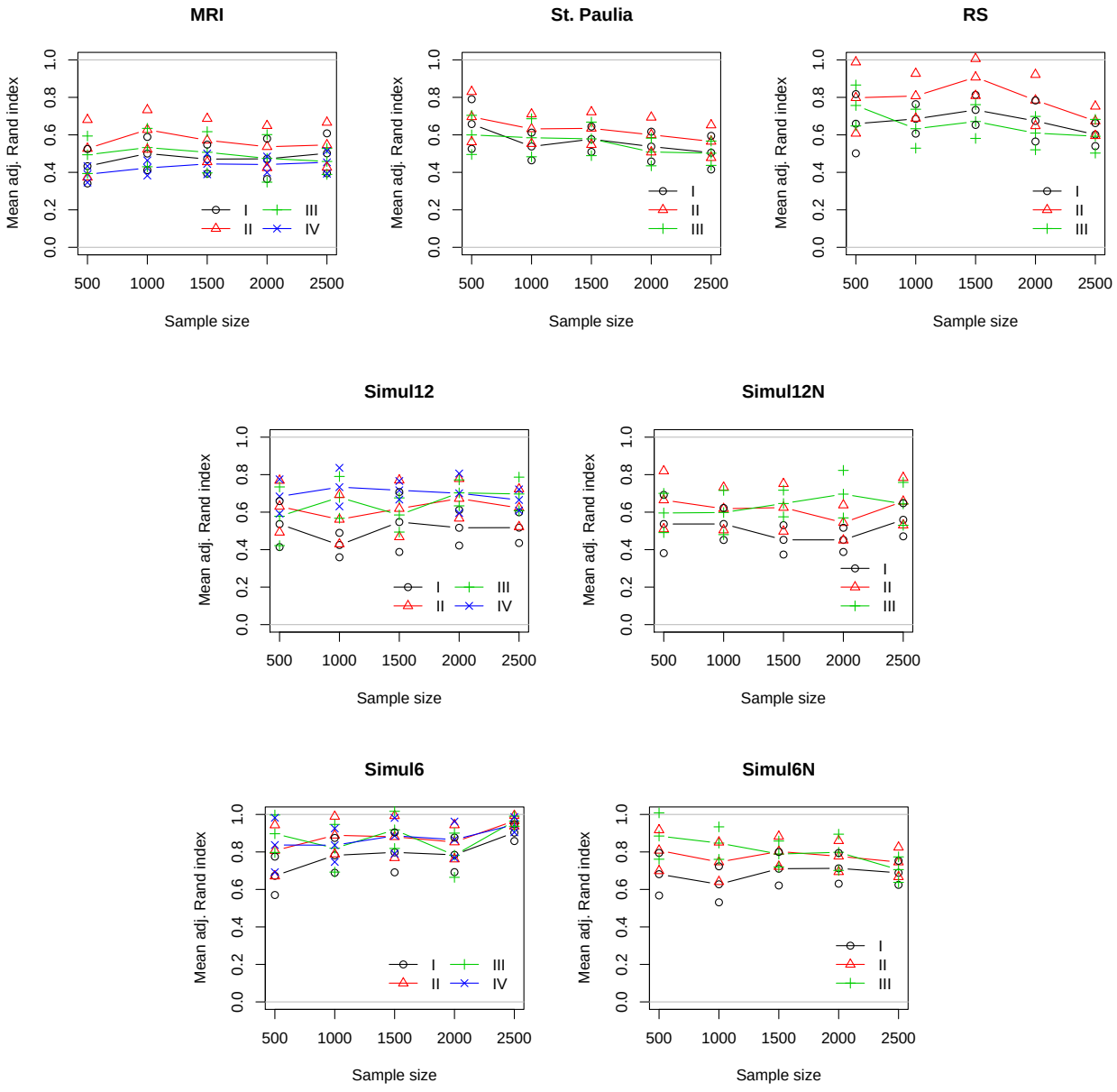

Figure 5: Agreements between clusterings from ten different samples, indicated by mean values of the adjusted Rand index. One standard deviation above and below the mean value are also shown.

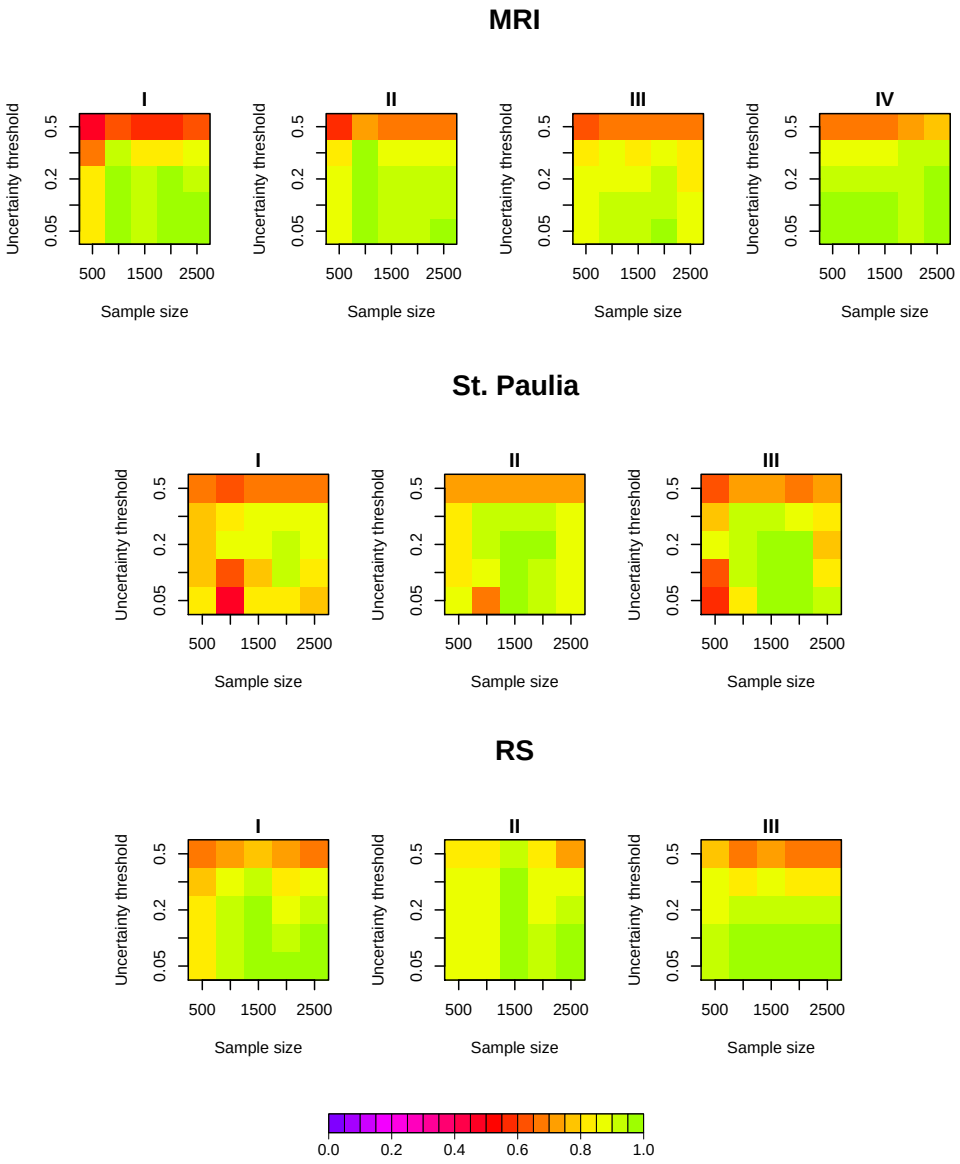

Figure 6: Classification agreements for the real data sets, as measured by mean adjusted Rand indices. The fraction of pixels taken into account is governed by the uncertainty of the classification (y-axis): the label 0.05, e.g., means that only pixels that are classified with an uncertainty smaller than 0.05 in all ten replicated segmentations are taken into account. Sample size is depicted on the x-axis.

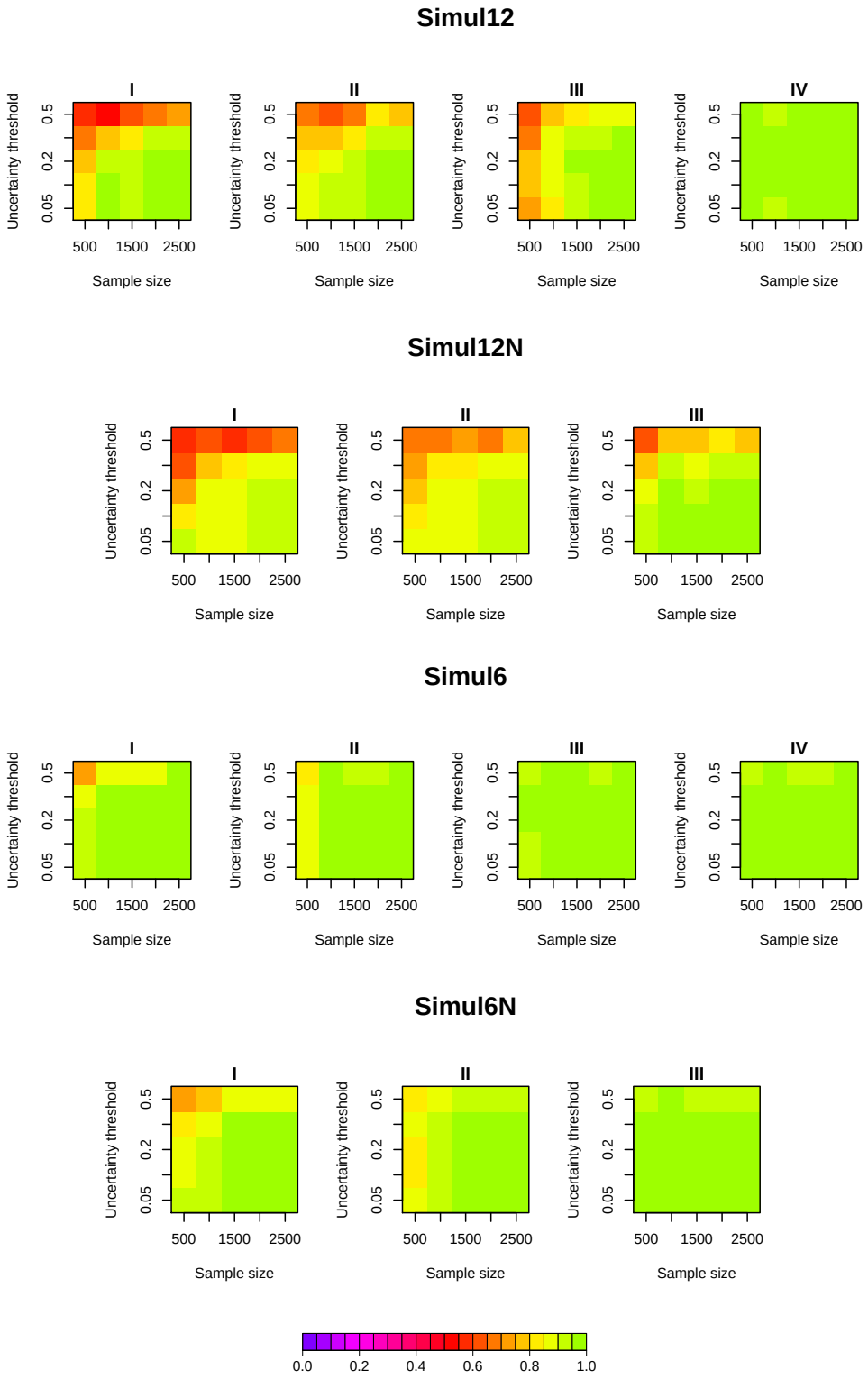

Figure 7: Classification agreements for the simulated data sets, as measured by mean adjusted Rand indices. See caption of Figure 6.

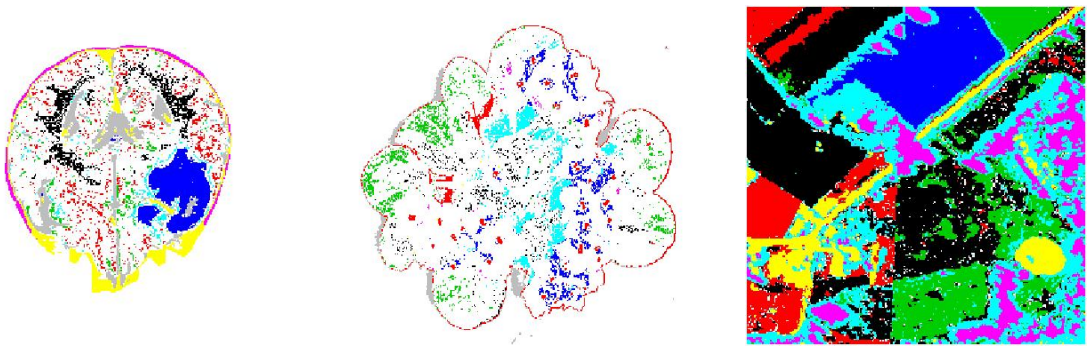

Figure 8: Stable classifications with adjusted Rand index greater than 0.9 (strategy II): for the three images, the uncertainty thresholds are .35, .2 and .5, respectively.

threshold we considered, 0.2, these numbers are 20, 10 and 60 percent. They do depend on the strategy employed and the sample size: typically, a larger sample, or a more complex strategy, will lead to more clusters and hence to larger uncertainty. Moreover, several clusters will disappear completely when our uncertainty threshold is severe. For the MRI and St. Paulia data sets, this happens for uncertainty thresholds of 0.2 or lower; the RS data set almost never loses complete clusters. As an example of which pixels are clustered with a low uncertainty, Figure 8 shows clusterings obtained with strategy II for the three real data sets, in a situation where the mean adjusted Rand index is at least 0.9: for the MRI image this corresponds to $u < 0.35$, for the St. Paulia image $u < 0.2$ and for the RS image $u < 0.5$. In the MRI image, the tumor and the cerebrospinal fluid clearly are stable clusters. The hearts of the flowers together with some remaining background, and shades on leaves and background form stable clusters in the St. Paulia image, and almost all pixels are part of a stable clustering in the RS image.

The loglikelihoods of the final segmentation for all sample sizes are summarized in Figure 9. Strategy IV leads to the largest likelihoods, strategy I the smallest. The difference between strategies I and II is completely due to the extra EM steps. The increase is larger with smaller samples; there is more room for improvement there. Also the differences between replicated runs are larger with smaller samples. In terms of higher likelihood, there seems to be little gain in having samples larger than 2000 objects.

For the simulated data, the likelihood of the "true" classification is known (indicated with a gray line in Figure 9). For the no-noise cases, all strategies seem to converge to these values with increasing sample size. The convergence is faster in Simul6, where fewer classes are present: there, all strategies except strategy I perform optimally for samples of size 1,000 or greater. In some of the noise cases, likelihoods higher than the likelihood with the true noise classification were found. This is to be expected: some noise points may lie very close to a "true" cluster and when classified to that cluster will lead to a higher likelihood.

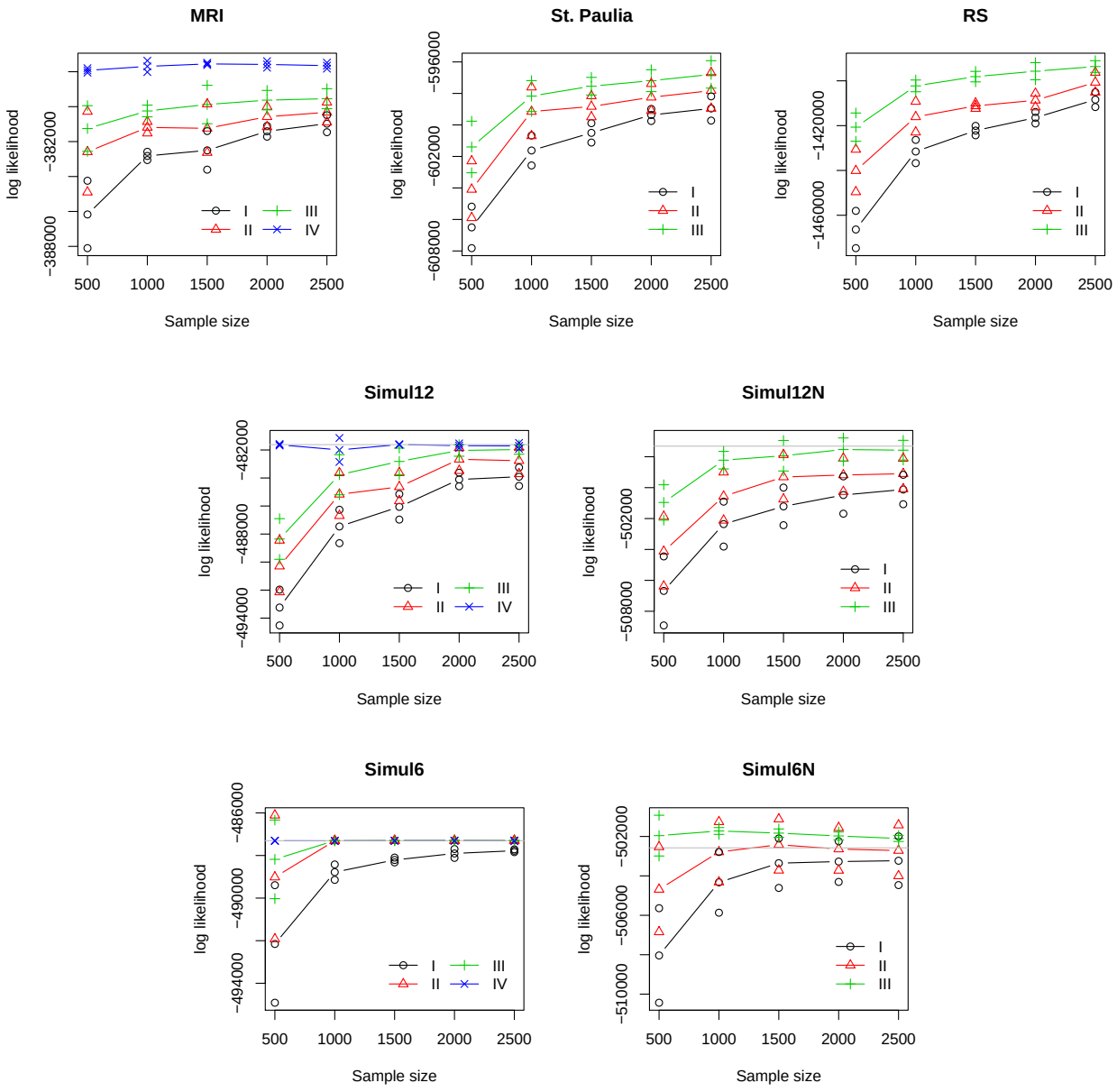

Figure 9: Loglikelihoods of the final model selections. Means are plotted for the ten repeated samples; plus or minus one standard deviation is given as well. The gray lines in the simulated data sets show the loglikelihood of the "true" solution.

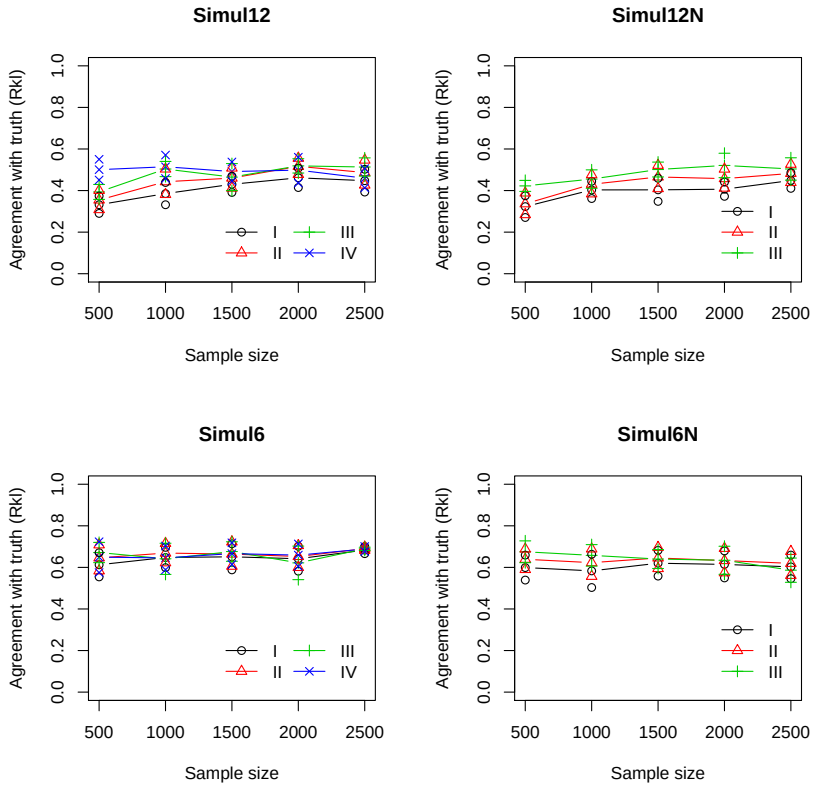

Figure 10: Agreement with "true" class labels for the simulated data sets (adjusted Rand index).

5.3 Accuracy

With the simulated data it is possible to assess the accuracy of the clusterings. The number of clusters as well as the cluster model are known, and cluster labels have been preserved. Measuring the agreement between the estimated and the "true" class by the adjusted Rand index, leads to the summaries depicted in Figure 10. The values are rather low, and more or less independent of sample size. More complex strategies do slightly better for small sample sizes, but the differences are small.

Again, it is the less certain pixels that cause these low agreements: distinguishing between pixels classified with different levels of uncertainty, it appears that the most certain pixels agree very well with the true classification (see Figure 11). For the Simul12 data set, strategy IV for all sample sizes leads to values of the adjusted Rand index higher than 0.9, provided the uncertainty is below 0.1. Strategy IV shows no dependence on sample size: the other strategies show better agreement when the sample size increases. Strategy III, with sample sizes equal to or larger than 2000, also reaches adjusted Rand indices above 0.9, but only for pixels with an uncertainty lower than 0.05. Strategy I at best reaches 0.85 with the largest sample and the smallest uncertainty. In contrast, in the less complex

Table 4: Approximate timings (MRI data set) for the different strategies (minutes, user time, Pentium III 1GHz processor). In this table, only the three most elaborate models (EEV, VEV and VVV) are considered with 4–14 clusters. The numbers cited are means of ten repeated clusterings.

	Strategy I	Strategy II	Strategy III	Strategy IV
N=500	0.4	0.6	1.6	23
N=1000	1.0	1.3	2.5	25
N=1500	1.7	2.0	3.7	25
N=2000	2.8	3.0	4.6	24
N=2500	4.1	4.2	5.8	24

Simul6 data set, all strategies achieve near-perfect matches. In this data set, the sample size does not matter much: only N=500 seems to be a bit too small for strategies I–III.

Adding noise generally decreases the agreement. Still, for the Simul6N data set, agreements better than 0.9 are obtained for all strategies at uncertainties smaller than 0.1; for strategies I and II, the sample size should be at least 1500 pixels. For strategy III, the sample size is not very important. The most difficult case is the Simul12N data set: mean agreements with the "true" classification are in all cases less than 0.9. The general trend, however, is clear again: larger samples show better agreements, and more expensive strategies (e.g. strategy III) are better than cheap ones (strategy I).

The main cause of the low agreements with the "true" values is not poor performance of the clustering algorithm, but, rather, the large overlap between clusters. This follows from the following experiment: if we take the "true" cluster parameters of the simulated data sets, and classify all points using one E-step, then the agreement with the "true" classification should form an upper bound of what can be achieved. Doing this, we find values that are almost equal to the results of clustering. As an example, for the Simul12 data set the adjusted Rand indices calculated in this way for pixels with uncertainties lower than 0.05, 0.1, 0.2, 0.35 and 0.5, are 0.96, 0.92, 0.82, 0.70 and 0.60, respectively, which agrees quite well with what can be achieved by the best strategies (strategy IV, and strategy III with a sample size of 2000 or larger).

5.4 Timings

To give a very rough indication of the computational efforts associated with the different strategies, Table 4 presents timings for the MRI data set. The timings given here are approximate and serve only as an indication; in particular, one should be aware that R is an interpreted language, the scripts used are not optimized for speed, input/output is performed, and the load of the computer may have varied from time to time.

Clearly, strategy IV is by far the most expensive strategy in terms of computing time. As expected, timings for strategy IV are not dependent on the sample size (and as we have seen, neither are the results). The differences in timing between the other three strategies

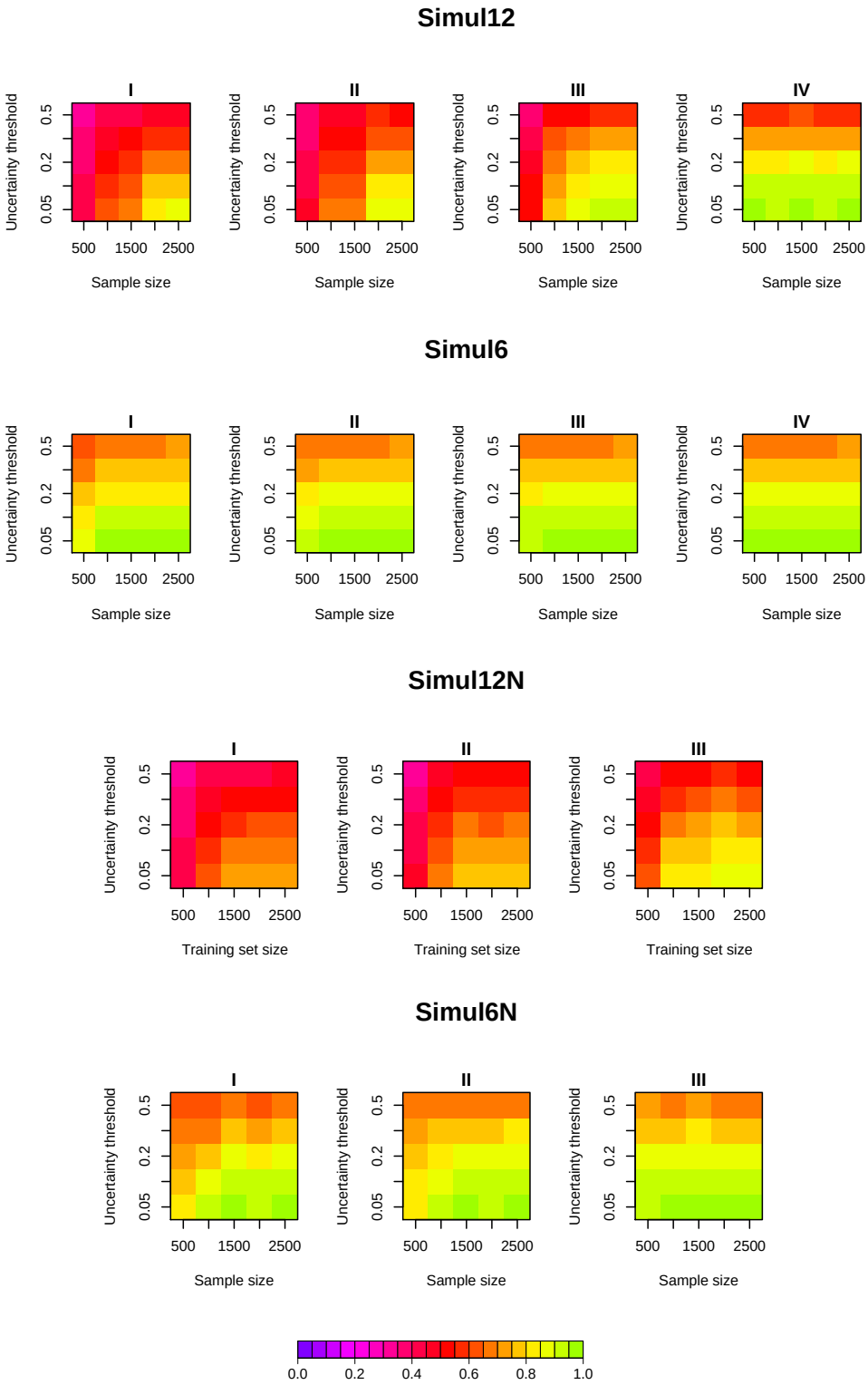

Figure 11: Agreement with "true" class labels (adjusted Rand index), dependent on the certainty of the classification.

are not very large: typically, strategy III takes only 1-2 minutes more than strategy I.

6 Discussion

We have proposed and experimented with several strategies for model-based clustering with large data sets by first applying model-based clustering to a sample of the data, and then extending the results to the full data set. There are large differences between the strategies employed. The simplest strategy, strategy I, performs worst. Overall, strategy III seems to strike a good balance between computing time and quality of results: for the data sets and computational setting considered here it takes only one or two minutes more than the fastest strategy, strategy I, but leads to higher likelihoods and better stability in classifications than the simpler strategies. Moreover, with strategy III, there is little advantage in having sample sizes larger than 2000 objects. Strategy IV, on the other hand, is considerably slower, regardless of sample size. Although in simulated cases the correct model was picked in essentially all cases, for the real data there was a tendency to pick a larger number of components, possibly caused by non-normality of the data.

Based on the results of the real and simulated data sets, strategy III with a sample size of 2000 seems to be the best of the approaches we have investigated: the loglikelihood of the final segmentation in this case was close to the "true" value, and agreements between different runs are also good, especially when looking at the more certain pixels. The gold standard, strategy IV, does not lead to much better results but takes at least five times as much computing time for the image data and simulations we considered.

Several extensions of the sampling approach are possible. First of all, one could use stratified sampling instead of random sampling. With images like the ones analysed here, one could focus on regions with a large variations, or the region of interest (e.g. the tumour region in the MRI image) and sample more densely in those areas. In cases like the RS image, one could decrease the number of pixels picked from a large, clearly recognisable field and concentrate more on regions with small objects. Although the samples would not be representative of the complete image, they may lead to better estimates of the number of clusters and cluster parameters.

Another potential improvement is not to rely on a single clustering, but to perform multiple clusterings from different random samplings, much like the setup of the experiments in this paper. If the best (e.g., based on the loglikelihood of the final segmentation) of five clusterings was picked, the conclusions presented here would still hold. In general, the improvement found with a more expensive strategy is larger than the improvement resulting from repeating the clustering several times with the same cheap strategy.

The one-step approach discussed here is attractive in its simplicity and speed of execution. Several other approaches to dealing with large data sets have been proposed. Fayyad and Smyth (1996) and Maitra (2001) propose forms of iterative sampling, where the second sample is taken from points that are poorly described by the clusters found so far. In other words, the points that are well described are removed from the data set and the remainder is clustered again. This may continue for several cycles; eventually a large number of

clusters is found. Several heuristic procedures are then used to merge clusters. Fraley and Raftery (2002) propose a similar idea, constructing the sample in the second iteration from a stratified sample of well-described points and poorly described points. This provides a way of avoiding the large numbers of clusters often found by the other iterative sampling methods.

Alternatively, the sample may be first partitioned into a large number of clusters using a simple clustering method, after which agglomerative hierarchical model-based clustering is initialized with these clusters rather than with the individual data points. Posse (2001) proposed a method based on the minimum spanning tree for obtaining the initial partition. A different approach called fractionation was proposed by Tantrum, Murua, and Stuetzle (2002) in the context of hierarchical model-based clustering, where the complete data set is split up randomly into smaller sets, which are clustered individually. The process is then repeated, but with the smaller sets formed by aggregated clusters rather than randomly.

Although these alternative methods may have an edge over the approach recommended in this paper when it comes to recognizing small clusters, any advantage would come at the expense of greater complexity (e.g. more user-defined parameters) and extra computation.

References

Banfield, J. and A. Raftery (1993). Model-based gaussian and non-gaussian clustering. *Biometrics 49*, 803–821.

Campbell, J., C. Fraley, F. Murtagh, and A. Raftery (1997). Linear flaw detection in woven textiles using model-based clustering. *Patt. Recogn. Lett. 18*, 1539–1548.

Campbell, J., C. Fraley, D. Stanford, F. Murtagh, and A. Raftery (1999). Model-based methods for textile fault detection. *Int. J. Imag. Sci. Techn. 10*, 339–346.

Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *Patt. Recogn. 28*(5), 781–793.

Dasgupta, A. and A. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *J.Am.Stat.Assoc 93*, 294–302.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B 39*(1), 1–38.

Fayyad, U. and P. Smyth (1996). From massive data sets to science catalogs: applications and challenges. In J. Kettenring and D. Pregibon (Eds.), *Statistics and massive data sets: report to the committee on applied and theoretical statistics*. National Research Council.

Fraley, C. (1998). Algorithms for model-based gaussian hierarchical clustering. *SIAM J. Sci. Comput. 20*, 270–281. Also: Technical Report no. 311, Department of Statistics, University of Washington.

Fraley, C. and A. Raftery (1998). How many clusters? which clustering method? - answers via model-based cluster analysis. *Computer J. 41*, 578–588.

Fraley, C. and A. Raftery (1999). MCLUST: Software for model-based cluster analysis. *J. Classif. 16*(2), 297–306. Also: Technical Report No. 342, Dept. of Statistics, University of Washington.

Fraley, C. and A. Raftery (2002a, October). MCLUST: software for model-based clustering, density estimation and discriminant analysis. Technical Report TR 415, Dept. of Statistics, University of Washington.

Fraley, C. and A. Raftery (2002b). Model-based clustering, discriminant analysis, and density estimation. *J.Am.Stat.Assoc 97*(458), 611–631.

Hubert, L. (1985). Comparing partitions. *J. Classif. 2*, 193–218.

Ihaka, R. and R. Gentleman (1996). R: A language for data analysis and graphics. *J. Comput. Graphic. Statist. 5*(3), 299–314.

Jain, A. and R. Dubes (1988). *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795. Also Technical Report no. 254, March 1993, Dept. of Statistics, University of Washington.

Kaufman, L. and P. Rousseeuw (1989). *Finding Groups in Data, An Introduction to Cluster Analysis*. New York: Wiley.

Maitra, R. (2001). Cluatering massive data sets with applications in software metrics and tomography. *Technometrics 43*(3), 336.

McLachlan, G. and K. Basford (1988). *Mixture models*. New York: Marcel Dekker.

McLachlan, G. and T. Krishnan (1997). *The EM algorithm and extensions*. Wiley.

McLachlan, G. and D. Peel (2000). *Finite mixture models*. New York: John Wiley & Sons.

McLachlan, G., D. Peel, K. Basford, and P. Adams (1999). The EMMIX software for the fitting of mixtures of normal and t-components. *J. Stat. Soft. 4*(2). On line publication: www.jstatsoft.org.

Mukherjee, S., E. Feigelson, G. Babu, F. Murtagh, C. Fraley, and A. Raftery (1998). Three types of gamma ray bursts. *Astroph. J. 508*, 314–327.

Posse, C. (2001). Hierarchical model-based clustering for large data sets. *J. Comp. Graph. Stat. 10*, 464–486.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc. 66*, 846–850.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist. 6*, 461–464.

Stanford, D. and A. Raftery (2000). Principal curve clustering with noise. *IEEE Trans. Pattern Anal. Mach. Intell. 22*, 601–609.

Stanford, D. and A. Raftery (2002). Approximate bayes factors for image segmentation: The pseudolikelihood information criterion (plic). *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 1517–1520.

Tantrum, J., A. Murua, and W. Stuetzle (2002, March). Hierarchical model-based clustering of large datasets through fractionation and refractionation. Technical Report 407, University of Washington, Seattle.

Wang, N. and A. Raftery (2002). Nearest neighbor variance estimation (nnve): Robust covariance estimation via nearest neighbor cleaning (with discussion). *J. Am. Stat. Assoc. 97*, 994–1019.

Wehrens, R., A. Simonetti, and L. Buydens (2002). Mixture modelling of medical magnetic resonance data. *J. Chemom. 16*, 1–10.

Yeung, K., C. Fraley, A. Murua, A. Raftery, and W. Ruzzo (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics 17*, 977–987.